Network Biology Approach to Complex Diseases

LECTURE 3.
Information flow

Teresa Przytycka

NIH / NLM / NCBI

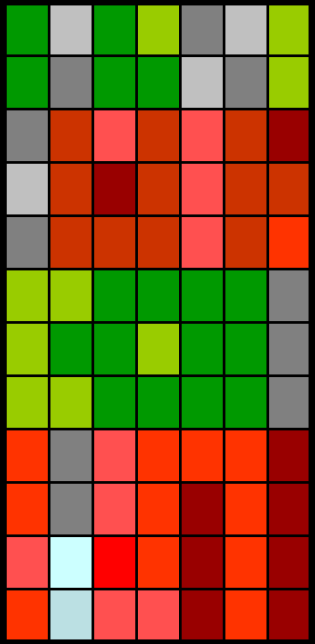
Recap from lectures 1-2

- We discussed approaches that use genotype and/or expression data to label genes as dys-regulated and search for modules containing such dys-regulated genes
- Some methods ensured additionally “consistency” of the modules (JACT, module cover)
- Emphasize of this lecture – information flow from genetic perturbations to gene expression perturbation

Information flow from genetoypc changes to expression changes

Copy number aberrations
or/and mutations

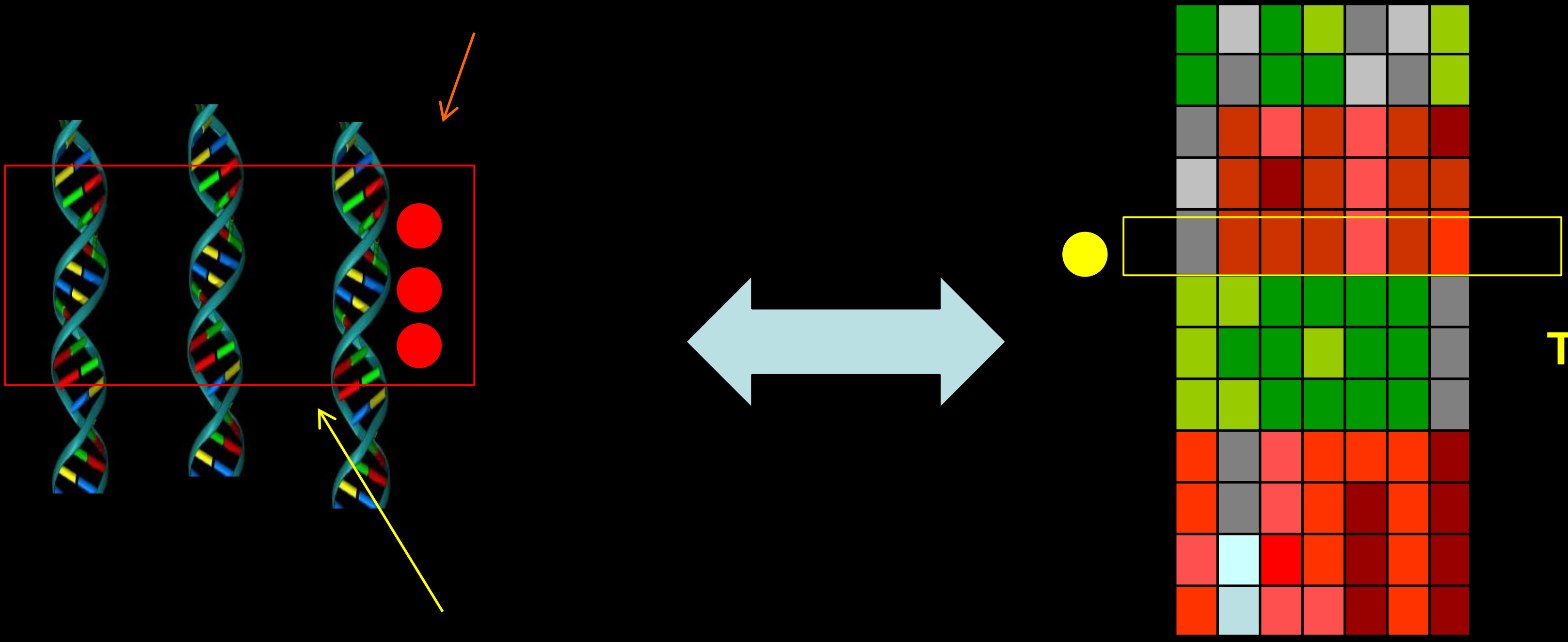
Gene
expression



Which gene is associated locus is most likely to drive the expression changes of a target gene?

Genes in the locus

Gene expression

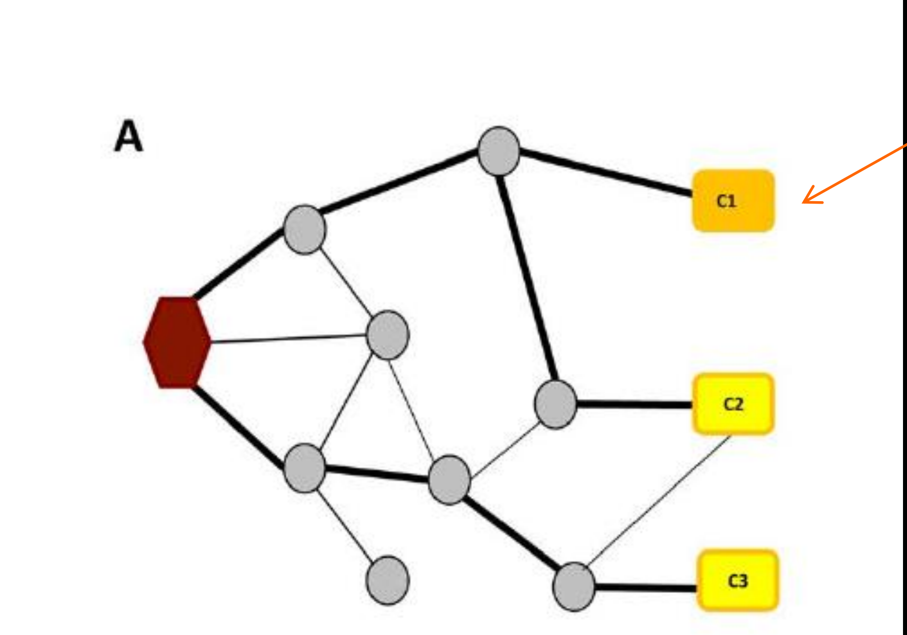


Target gene

Locus with genotypic changes that correlate with expression changes of target gene

Kim et al. PloS CB 2011

Shortest path approach



Target gene

Gene predicted to be most likely cause

Possible causal genes

Assumptions

- the gene closest in the network is the most likely driver
- genes on the shortest path are the intermediate nodes

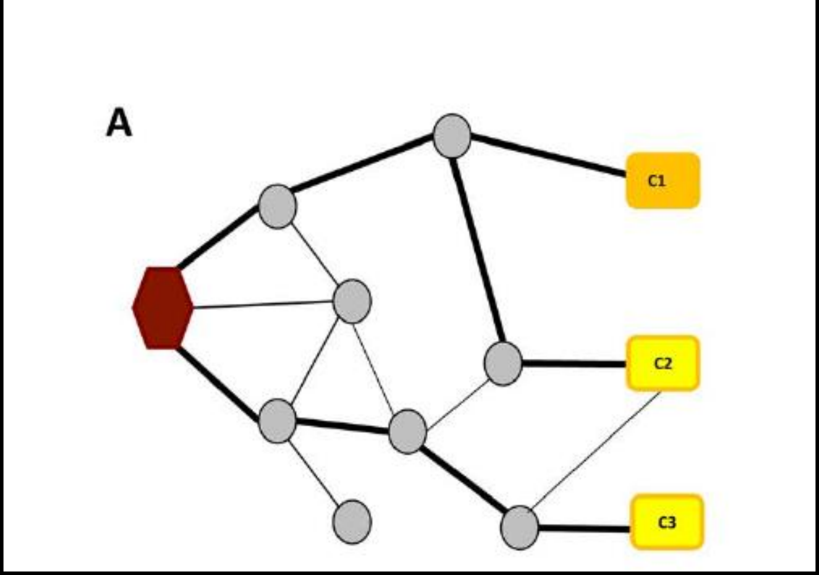
Advantages: the simplest assumption one can make in absence of additional information

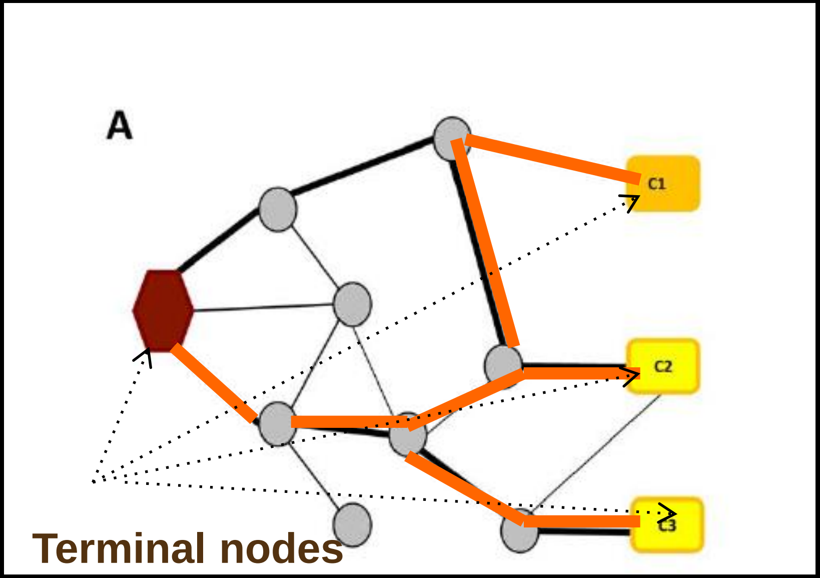
Disadvantages: Does not utilize expression data; Strongly impacted by netwokr bias and noise

Steiner tree

- Analogous to the shortest path idea: find a minimum size tree connecting all selected nodes

(thus individual paths might not be shortest possible but rather the total is minimized)




Steiner tree

Context - we assume C1,C2,C3 influence the target node and we use Steiner tree to model how information is propagated

Comment - many equivalent solutions might exist

Example

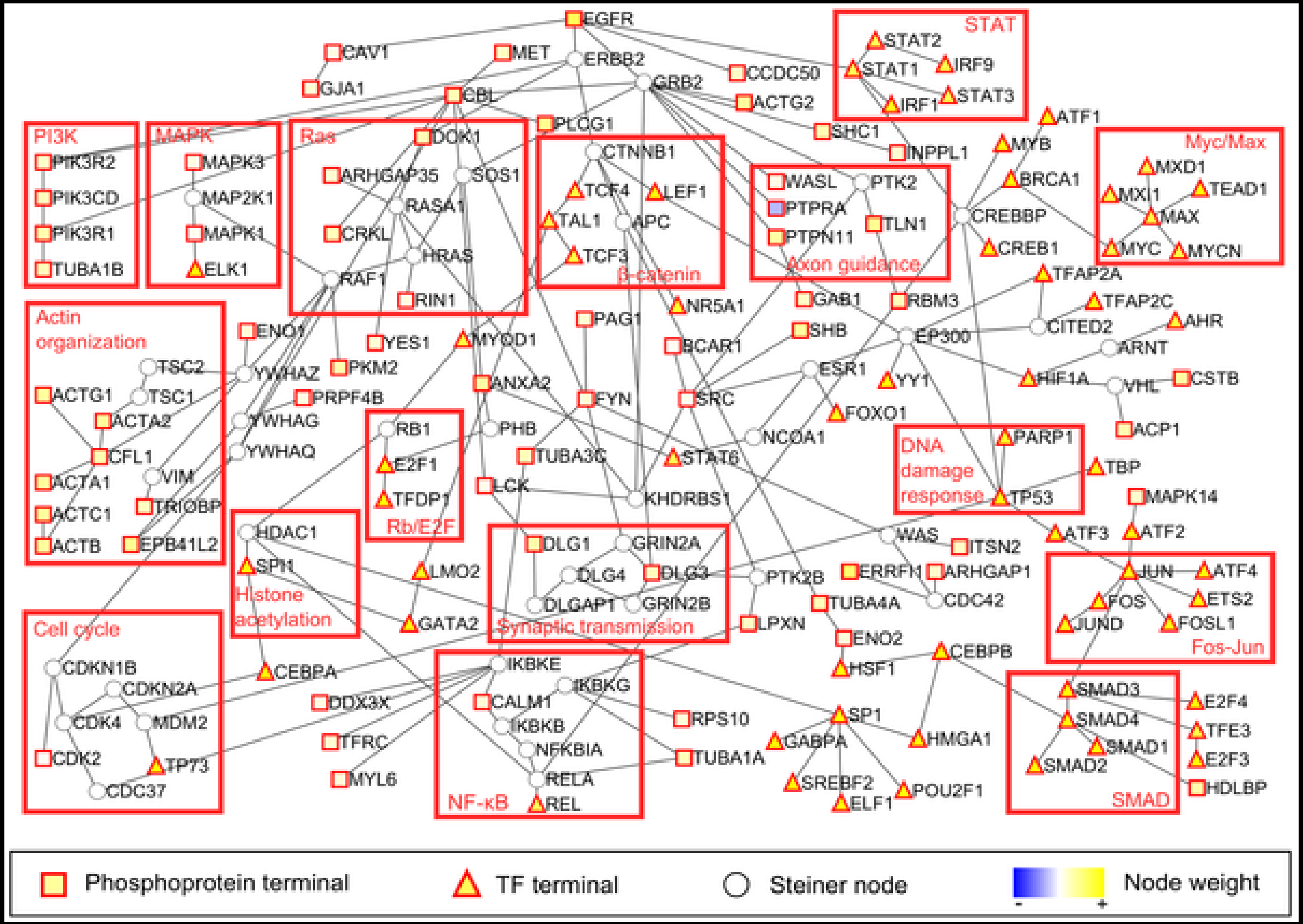
Huang S.S., Fraenkel E. Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks *Science Signaling* 2(81):ra40

Prize-collecting Steiner tree problem where not all the termini are required to be included in the solution.

- There is a cost of not including a terminal node
- There is a price for using edges to include a terminal in the network.
- Find minimum-weighted subtree that connects a subset of the termini to each other through the edges of the interactome graph and additional nodes not in the terminal set

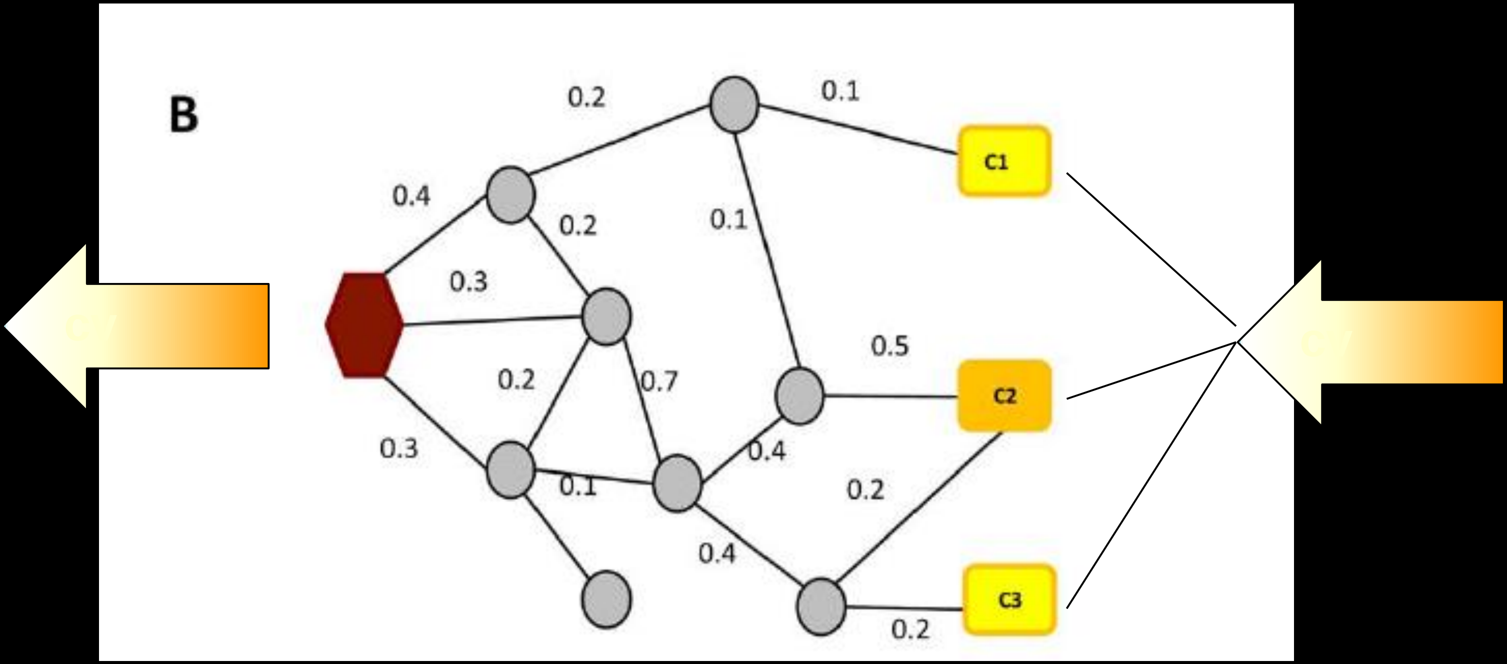
In Huang et al, a parameter β weights the penalties of excluding terminal nodes relative to the cost of including edges

Results using a variant of the method integrating optimal and suboptimal Steiner trees terminal nodes for comparative analysis two glioblastoma cell lines with different expression of EGFRvIII)



Huang S-sC, Clarke DC, Gosline SJC, Labadorf A, et al. (2013) Linking Proteomic and Transcriptional Data through the Interactome and Epigenome Reveals a Map of Oncogene-induced Signaling. PLoS Comput Biol 9(2): e1002887. doi:10.1371/journal.pcbi.1002887
http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1002887

Flow based approaches



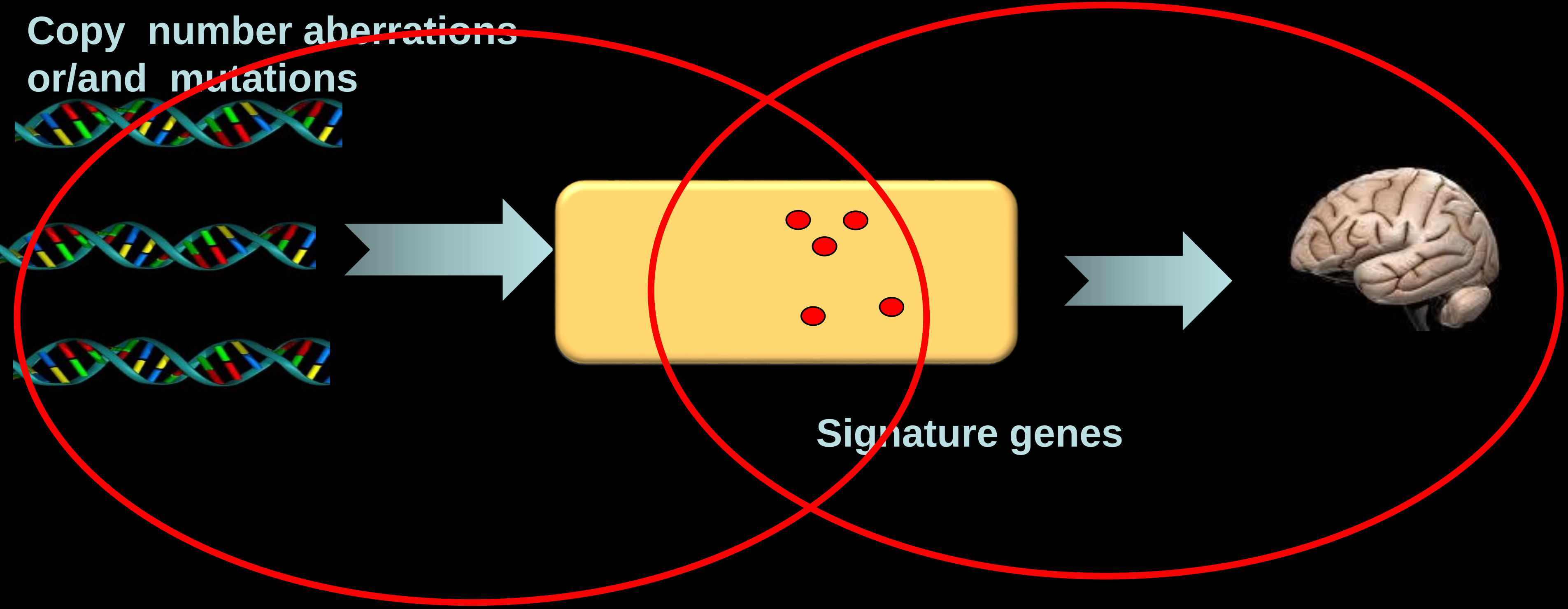
Current Flow - edges have resistance
Network Flow - edges have capacitances

Key Component: Kirchhoff low or flux balance requirement

eQTLNet

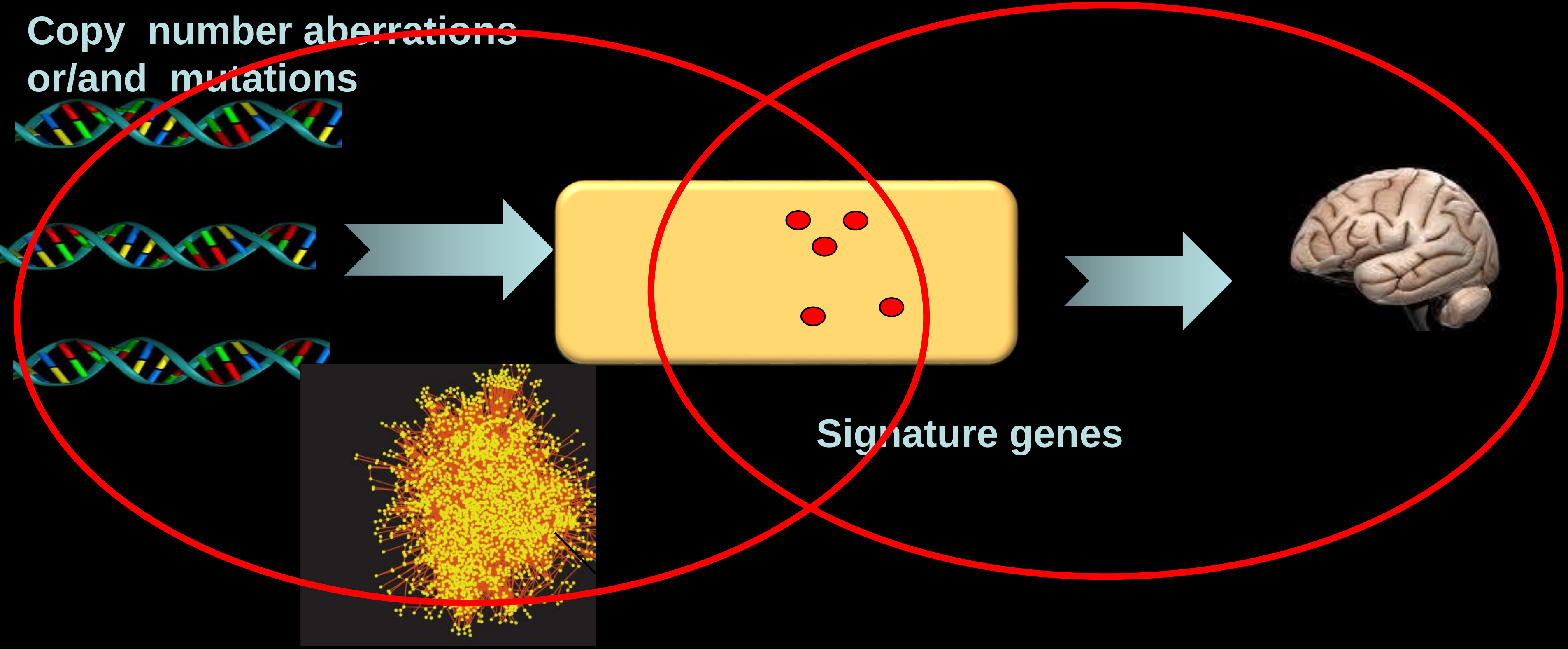
Combines eQTL analysis with network information and network flow approaches

Kim et al. PloS CB 2011



Copy number aberrations or/and mutations

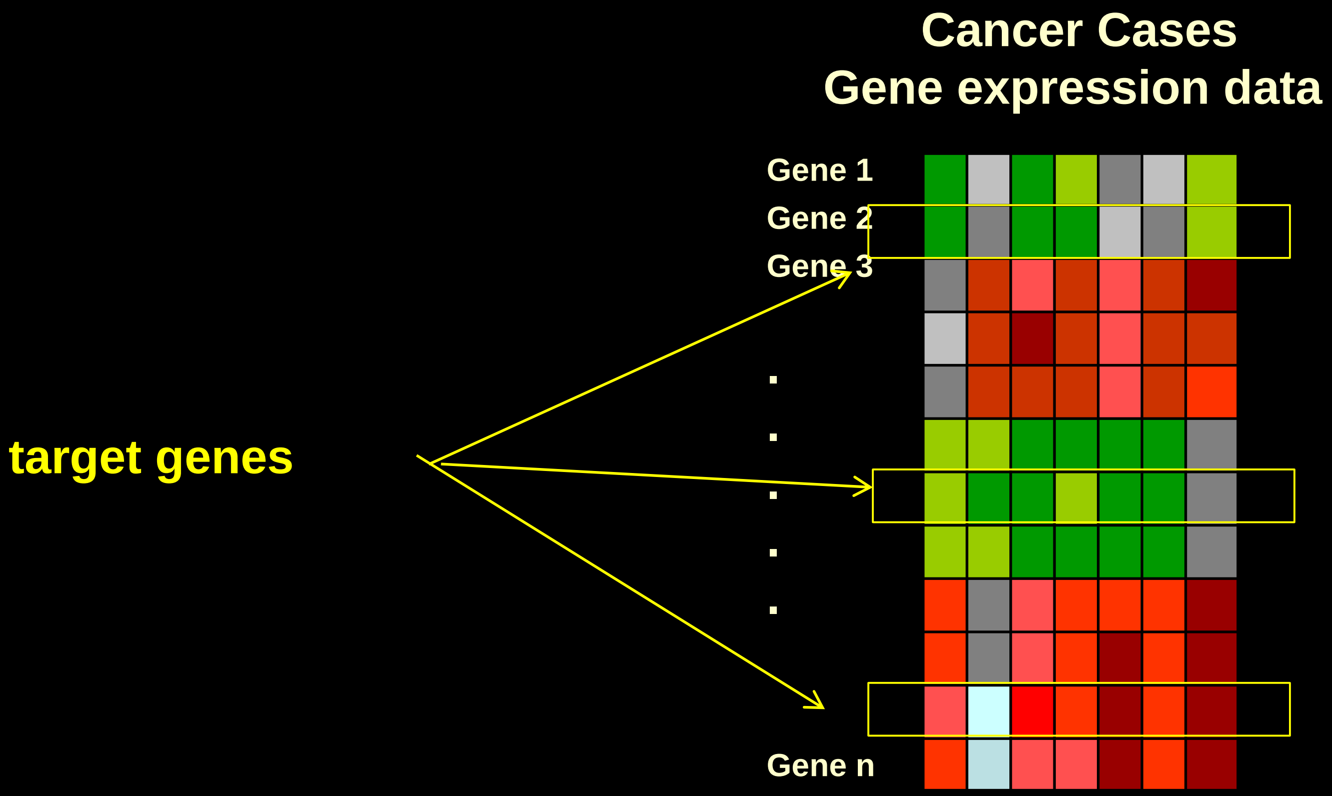
Signature genes



Copy number aberrations
or/and mutations

Signature genes

Selecting “signature” genes



target genes

Cancer Cases
Gene expression data

Gene 1

Gene 2

Gene 3

.

.

.

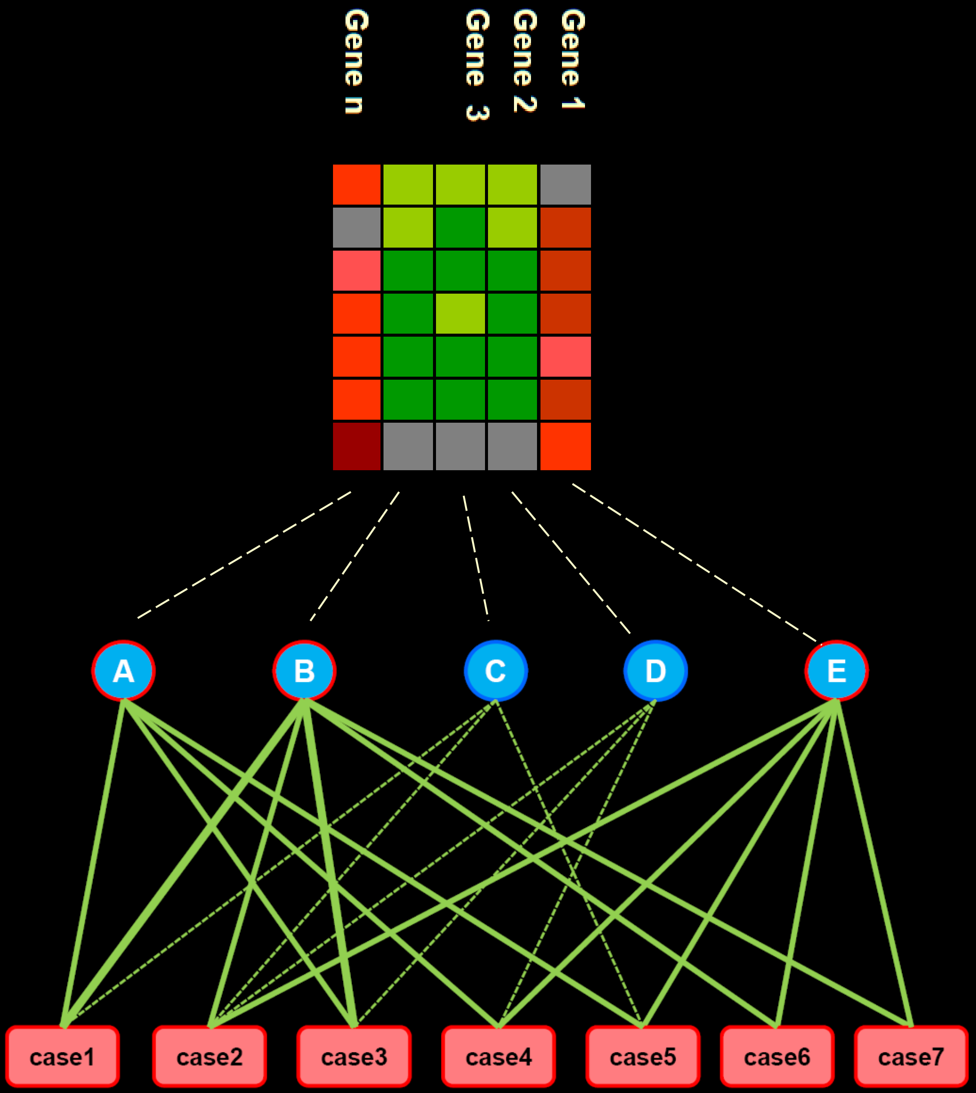
.

.

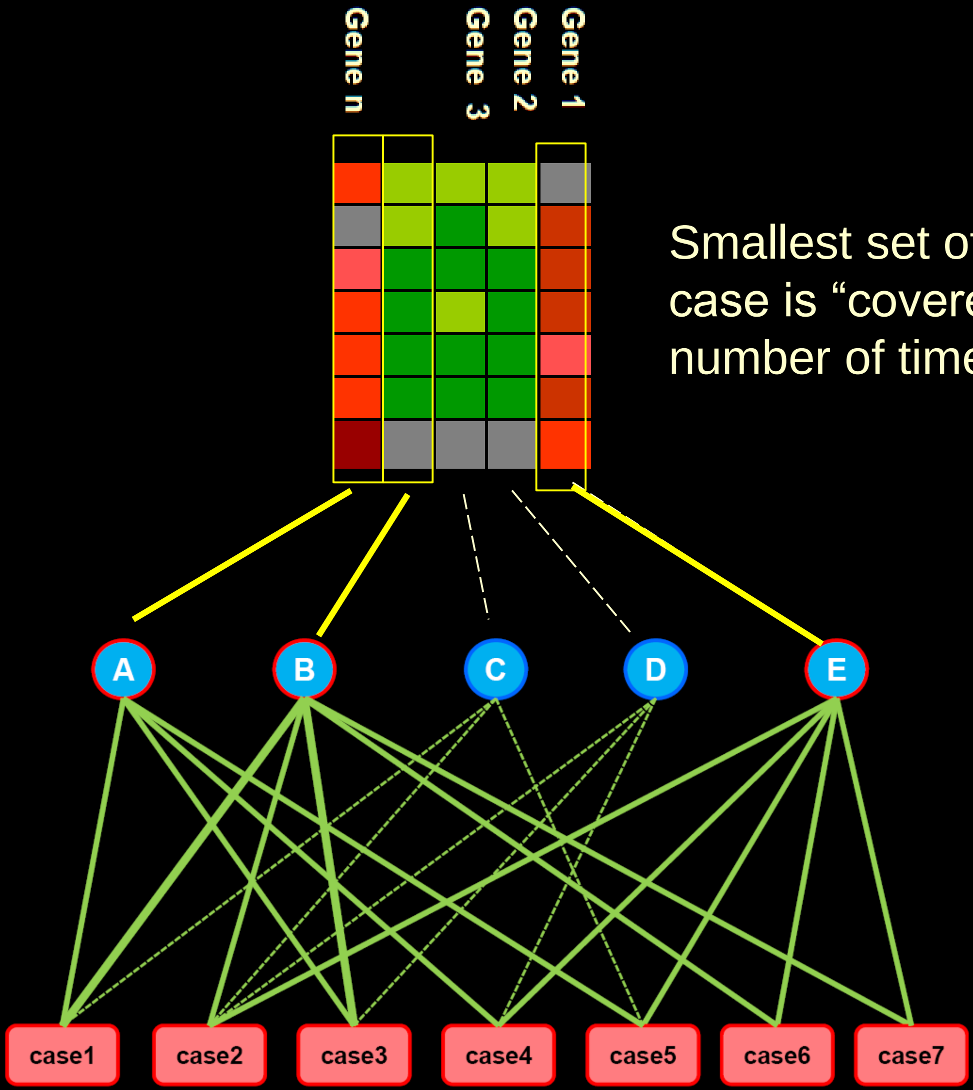
Gene n

Kim et al. PloS CB 2011

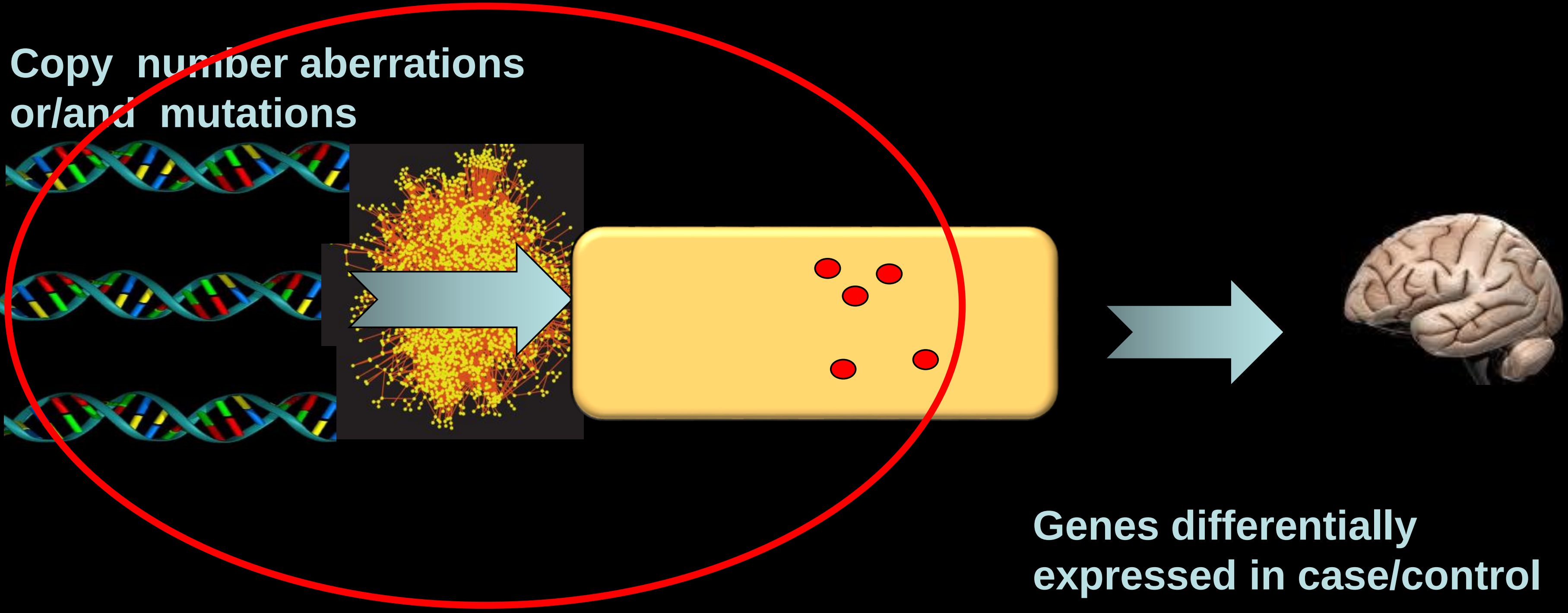
Selecting “signature” genes



Selecting “signature” genes



Smallest set of genes so that each case is “covered” at least specified number of times

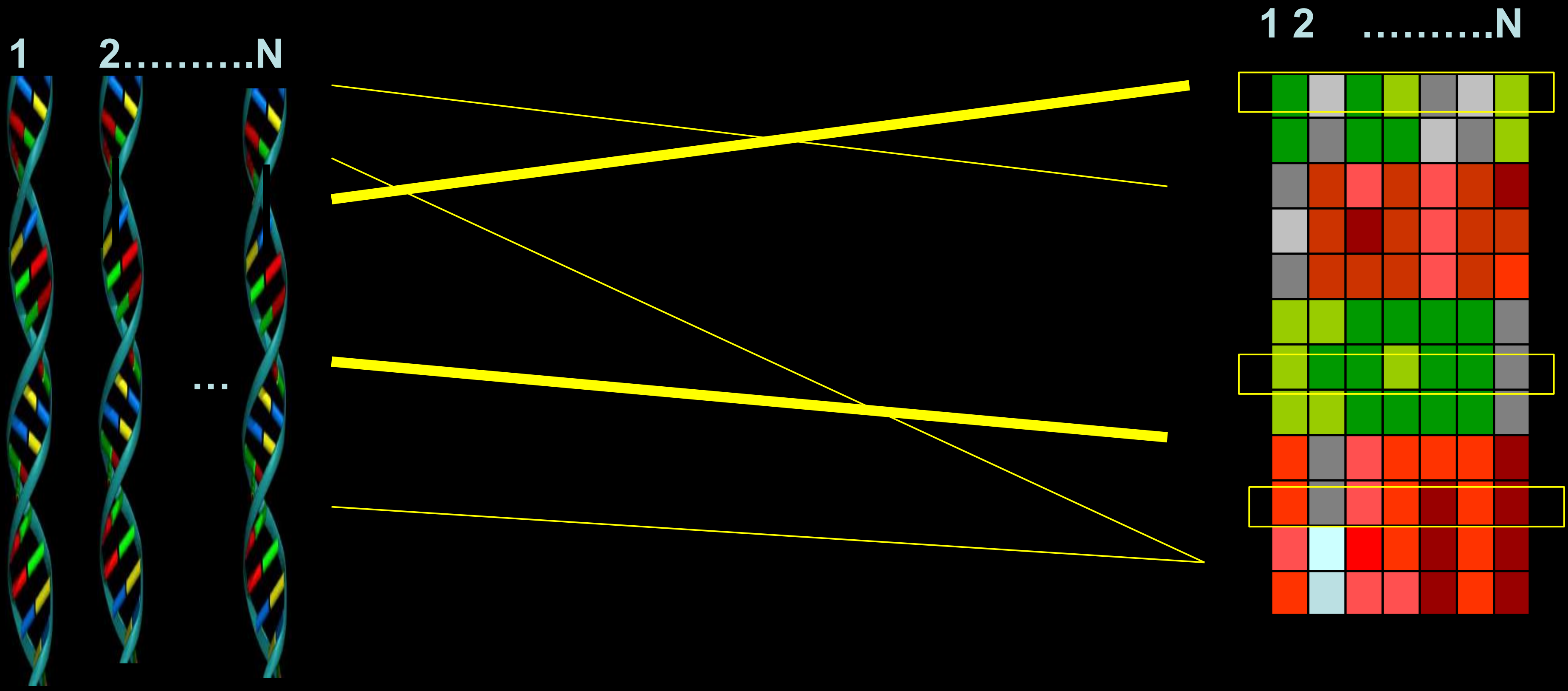


Copy number aberrations or/and mutations

Genes differentially expressed in case/control

Kim et al. PloS CB 2011

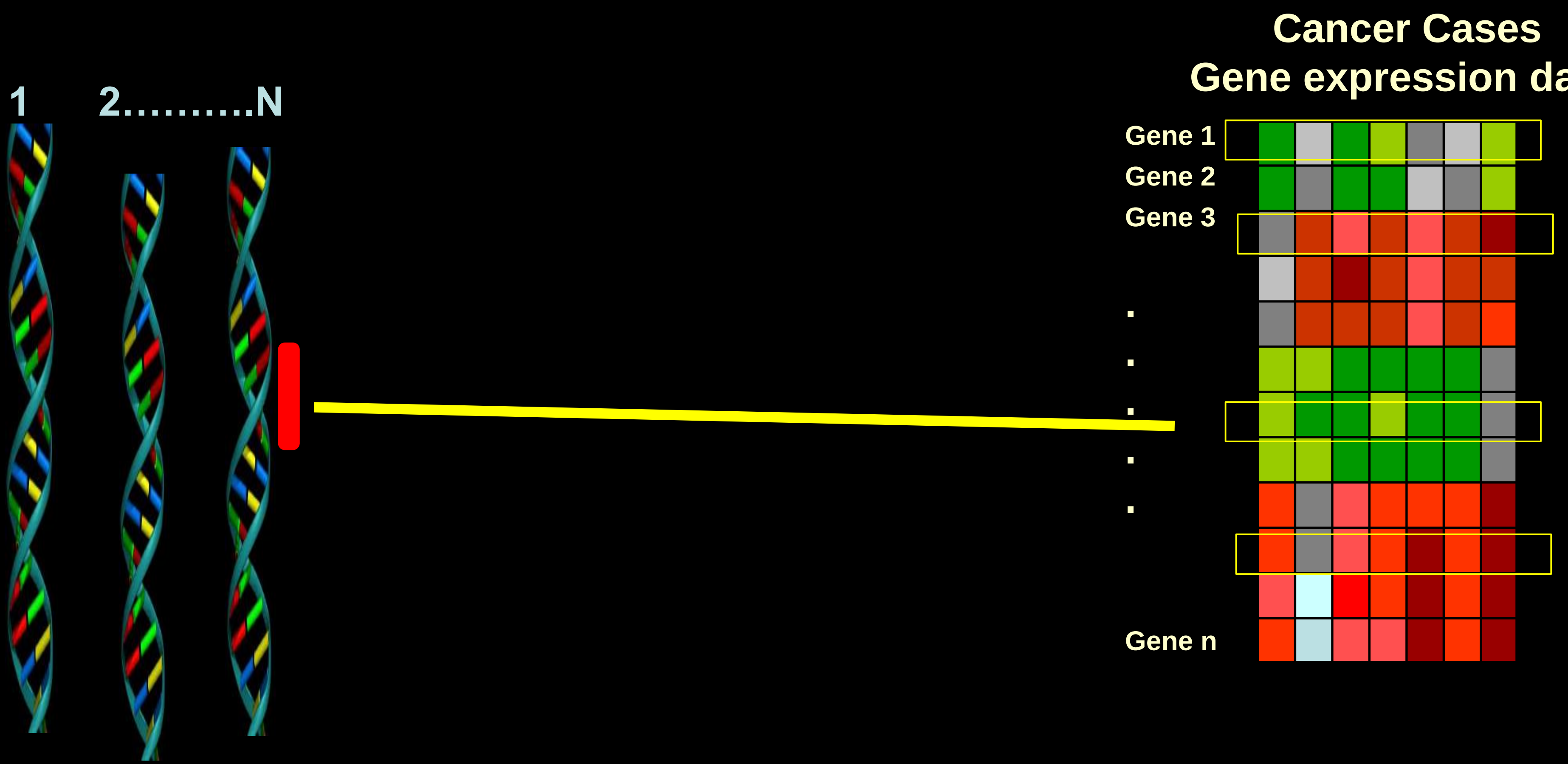
Associations between copy number variations and gene expression of selected target genes



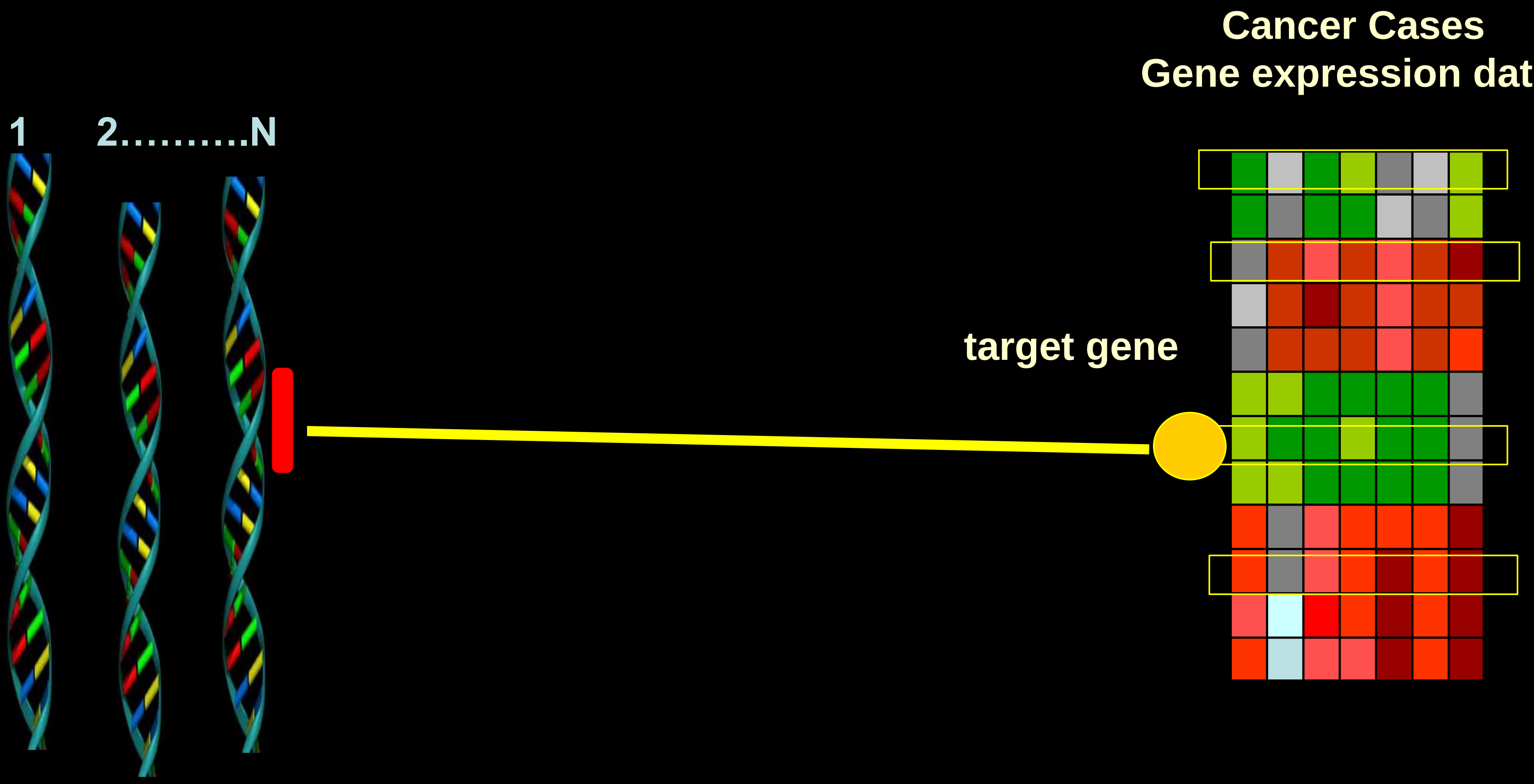
Cancer Cases
CNV data

Cancer Cases
Gene expression data

Significant correlation between CNV and expression



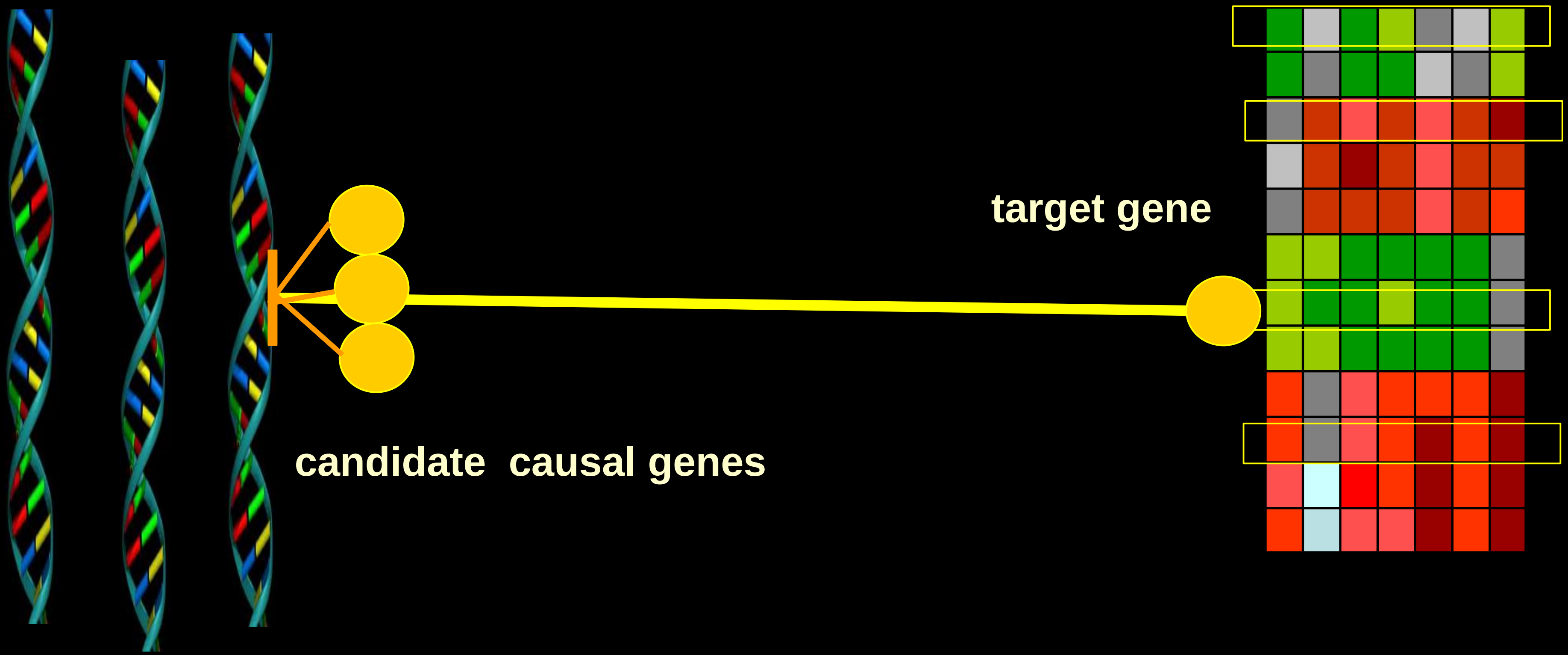
Significant correlation between CNV and expression



Cancer Cases
Gene expression dat

eQED idea from Ideker group

Significant correlation between CNV and expression



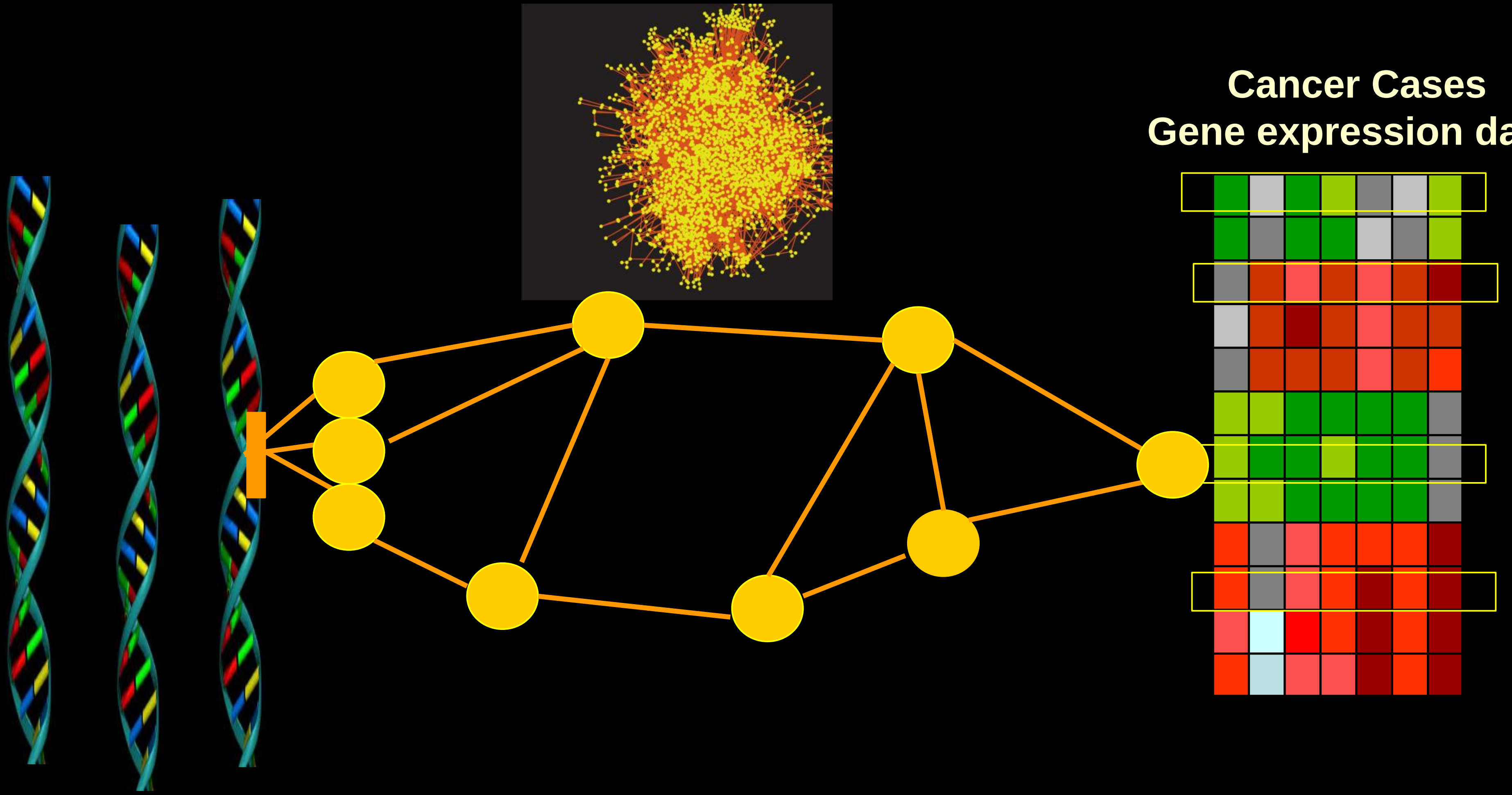
Cancer Cases
Gene expression da

target gene

candidate causal genes

eQED idea from Ideker group

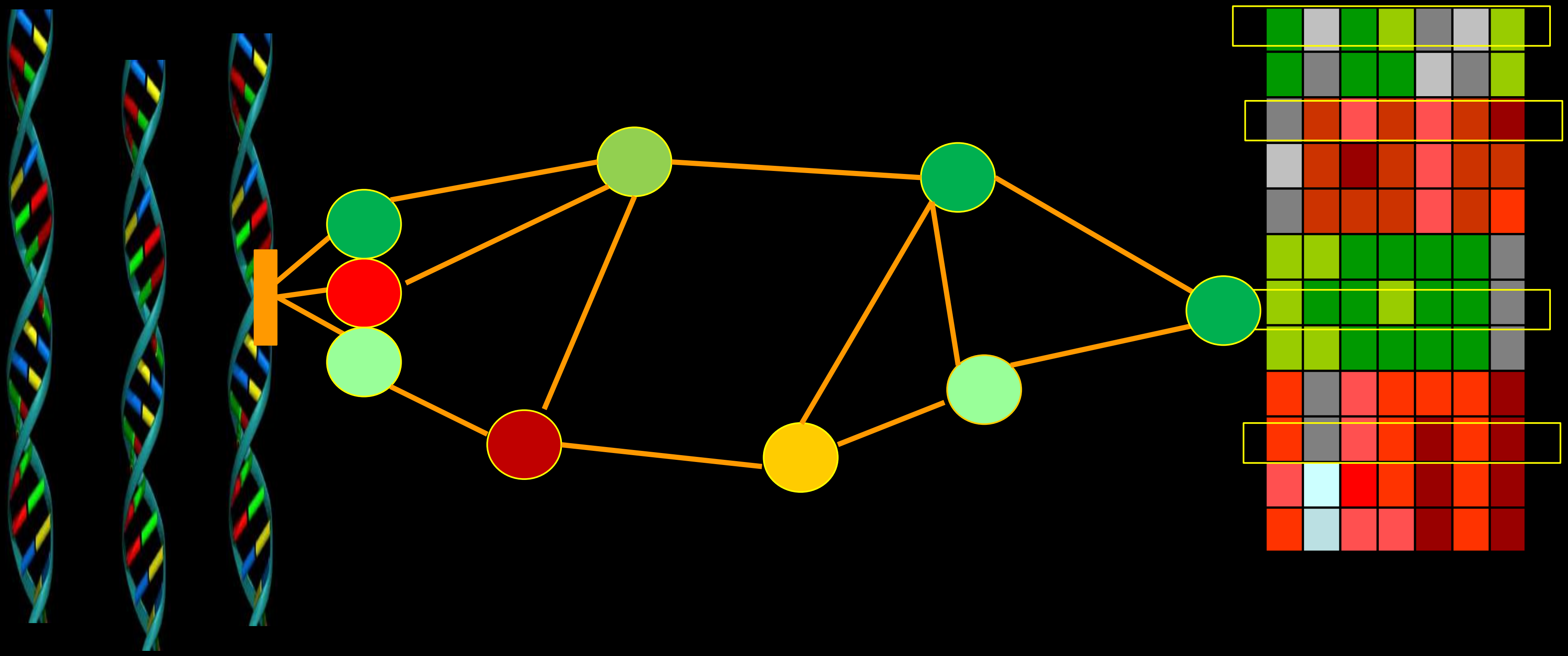
Uncovering pathways of information flow between CNV and target gene



Cancer Cases
Gene expression da

Kim et al. PloS CB 2011

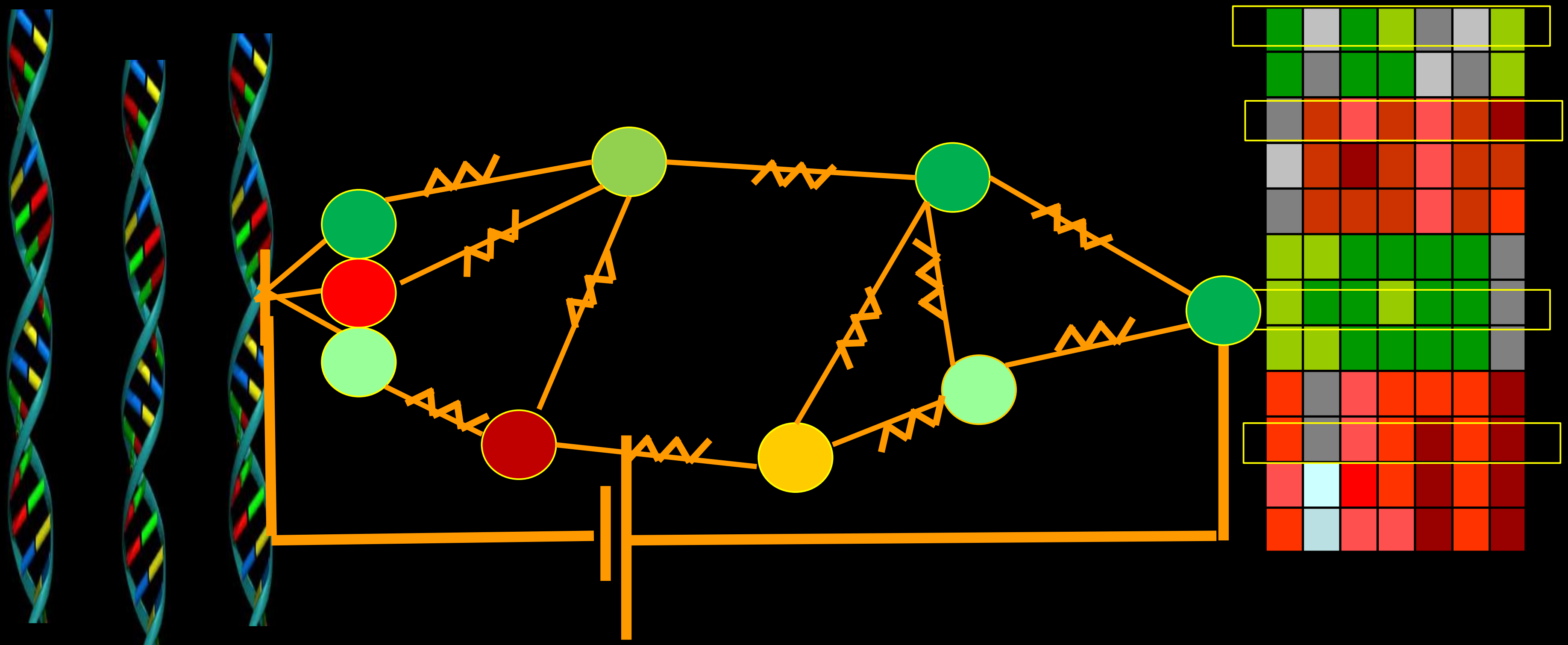
Using expression to guide path discovery



Cancer Cases
Gene expression da

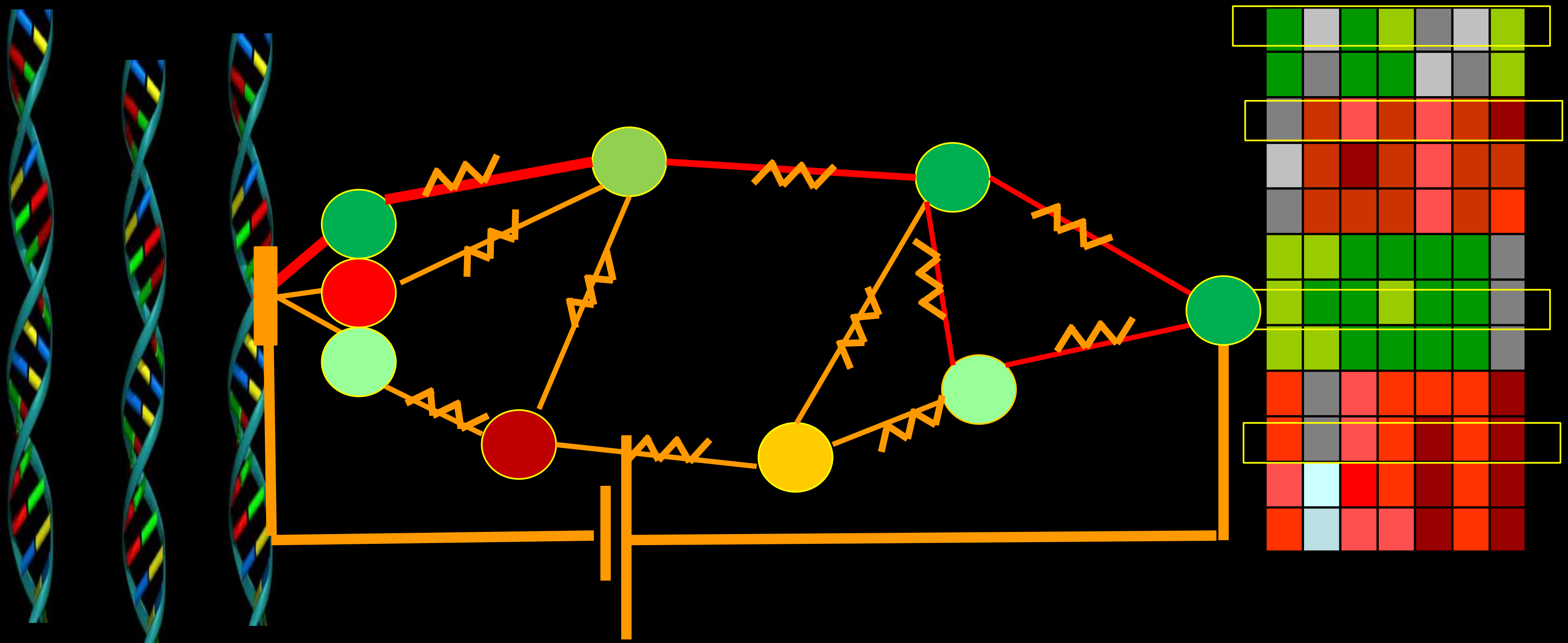
Kim et al. PloS CB 2011

Translating probabilities it resistances



Resistance **- set to favor most likely path -based on gene expression values**
(reversely proportional to the average correlation of the expression of the adjacent genes with expression of the target gene)

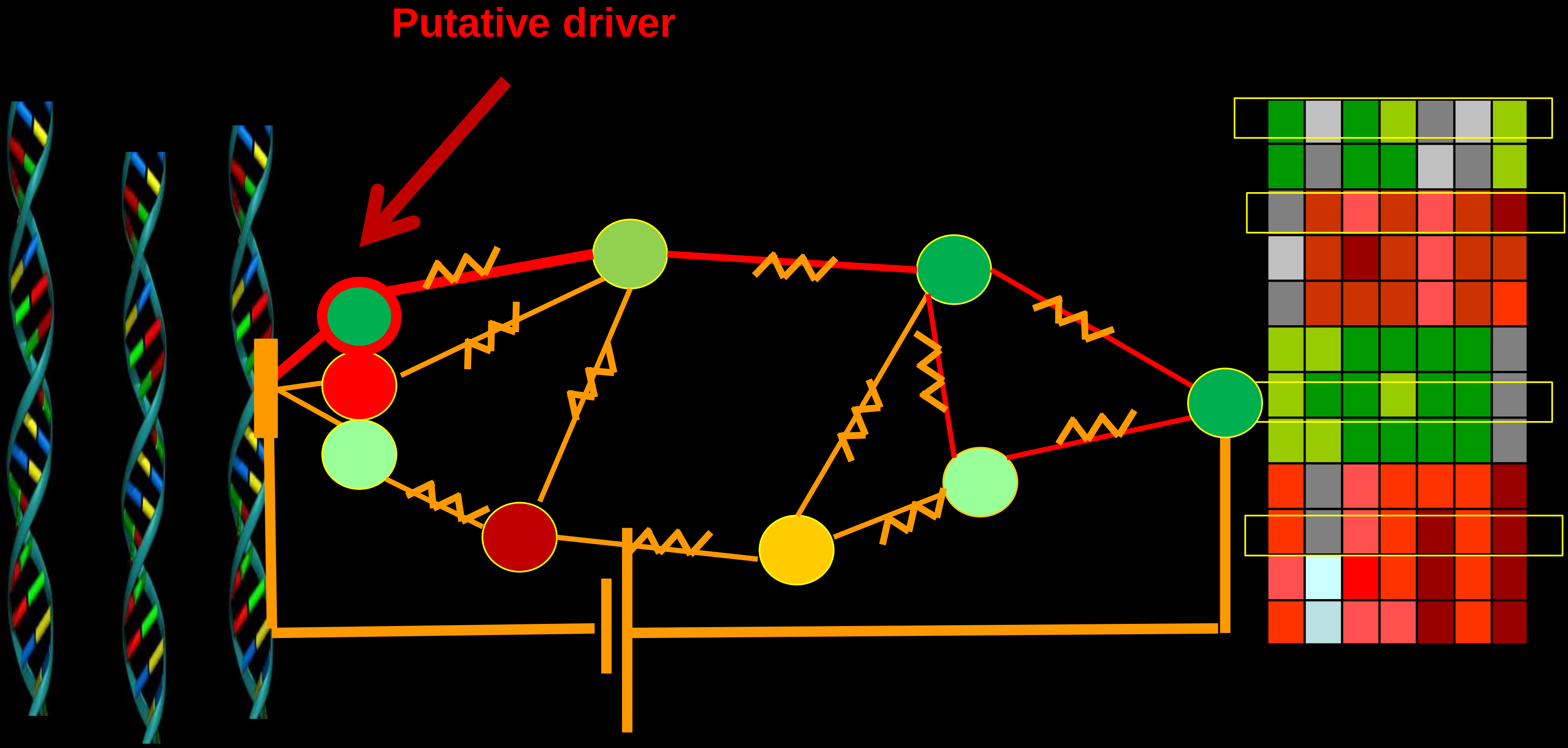
Finding subnetworks with significant current flow



Resistance **- set to favor most likely path -based on gene expression values**

(reversely proportional to the average correlation of the expression of the adjacent genes with expression of the target gene)

Finding subnetworks with significant current flow

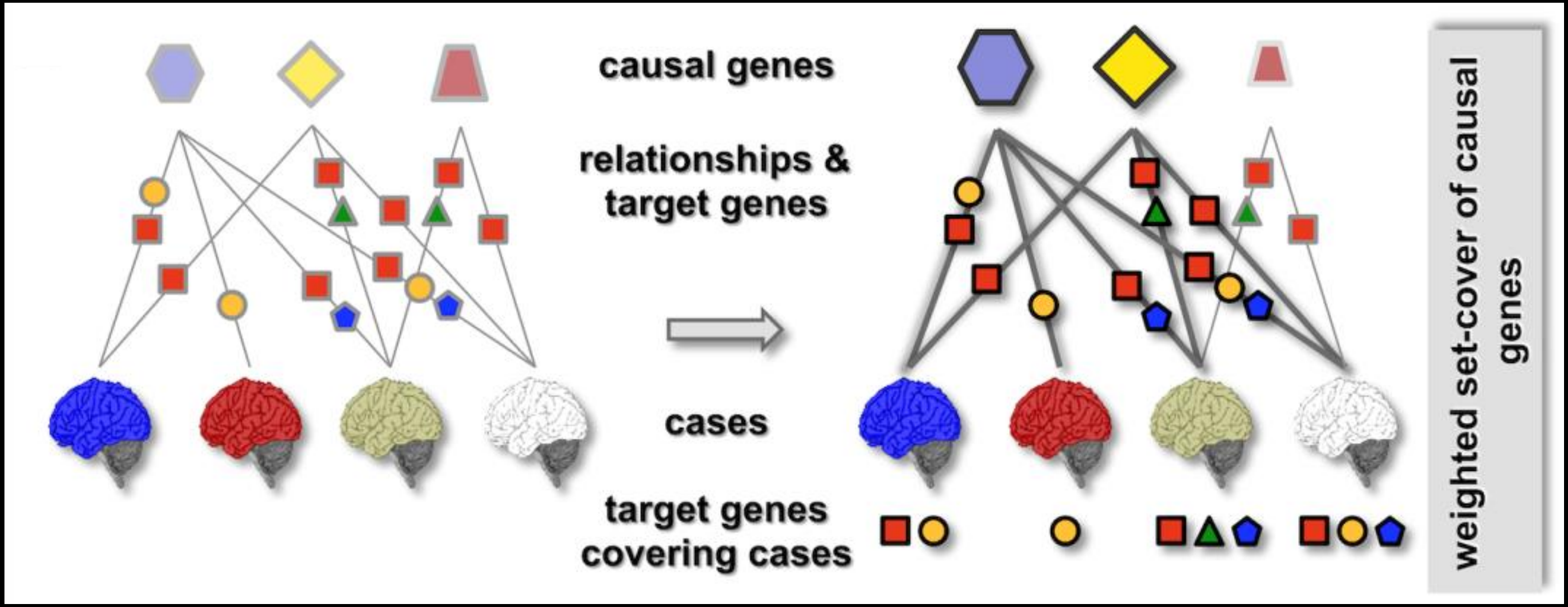


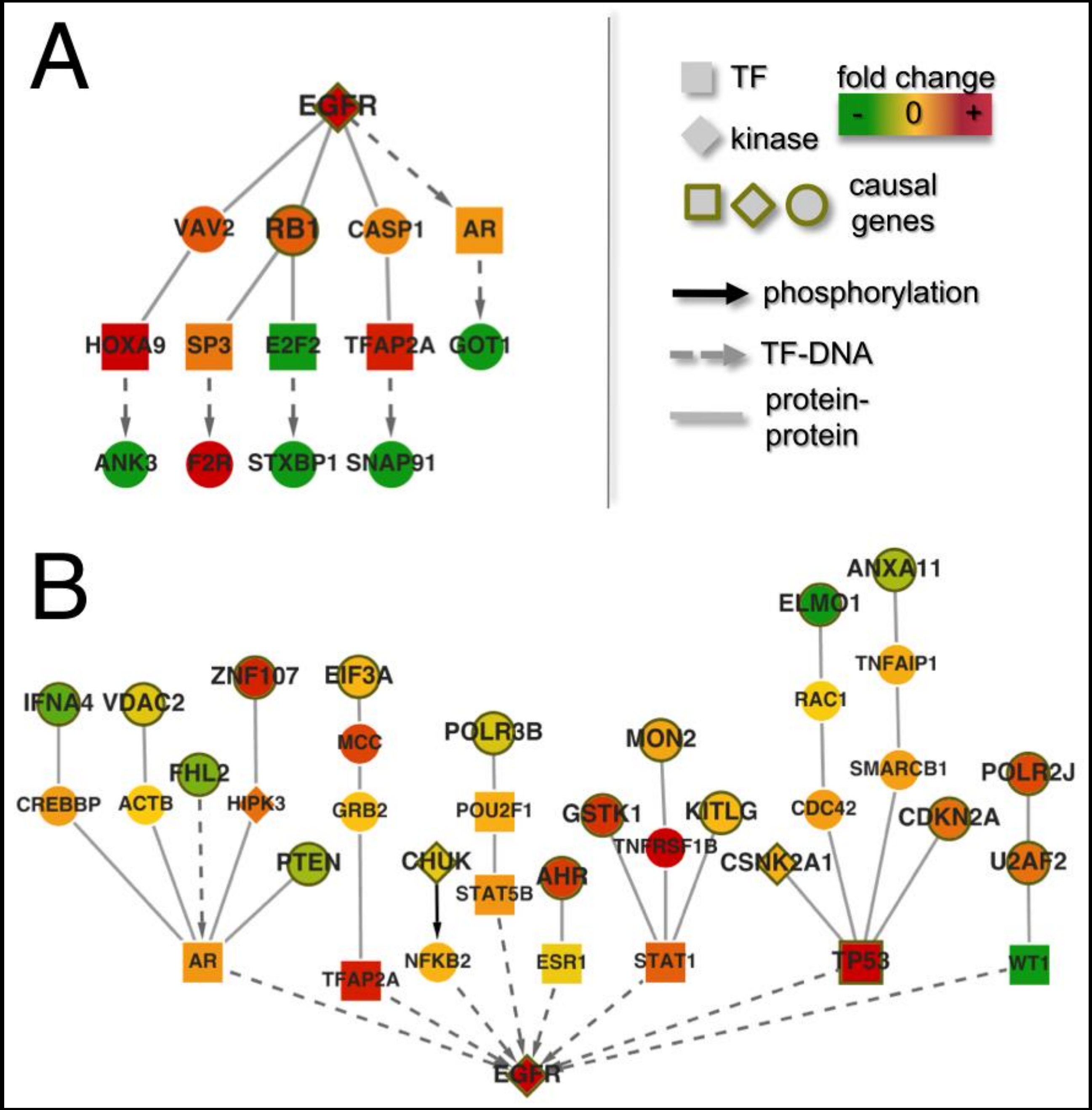
Resistance **- set to favor most likely path -based on gene expression values**
(reversely proportional to the average correlation of the expression of the adjacent genes with expression of the target gene)

Selecting causal genes

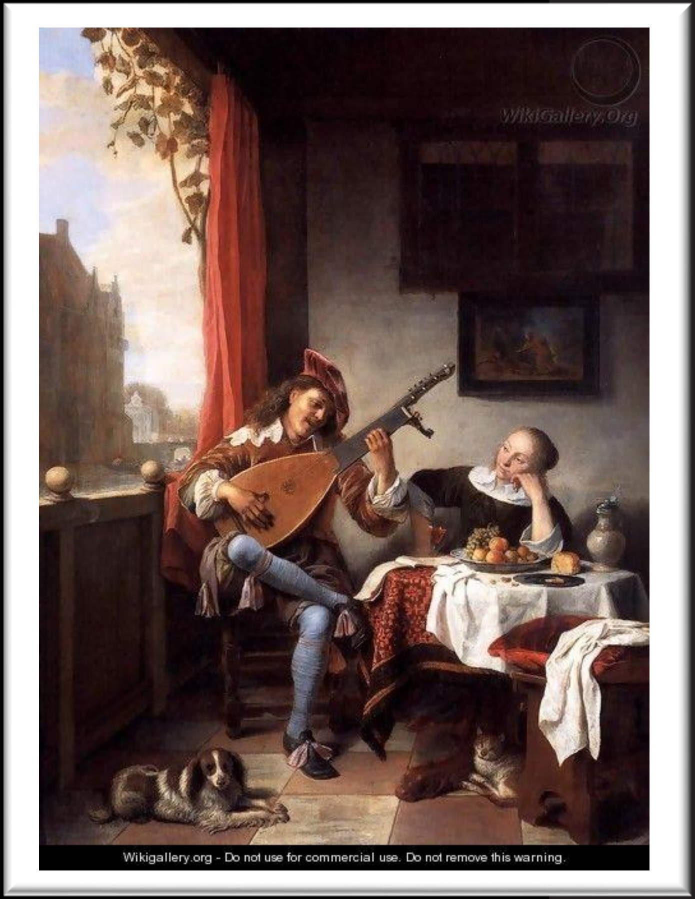
(weighted vertex cover)

Causal gene has copy number variation in the given case and low p-value pathway connecting it to a target gene that is differentially expressed in the same case; # of such target genes = edge weight

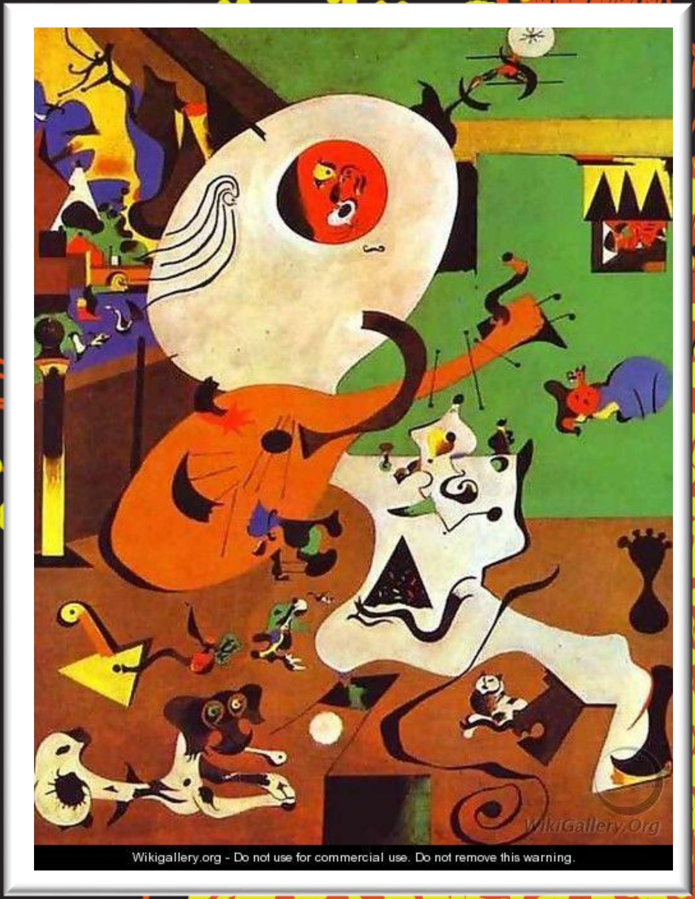

causal genes
relationships & target genes
cases
target genes covering cases
weighted set-cover of causal genes


A
EGFR
VAV2
RB1
CASP1
AR
HOXA9
SP3
E2F2
TFAP2A
GOT1
ANK3
F2R
STXBP1
SNAP91
TF
kinase
fold change
-
0
+
causal genes
phosphorylation
TF-DNA
protein-protein
B
ANXA11
ELMO1
TNFAIP1
RAC1
IFNA4
VDAC2
ZNF107
EIF3A
MCC
POLR3B
MON2
SMARCB1
POLR2J
FHL2
CREBBP
ACTB
HIPK3
GRB2
POU2F1
GSTK1
KITLG
CDC42
CDKN2A
TNFRSF1B
PTEN
CHUK
STAT5B
AHR
CSNK2A1
U2AF2
AR
TFAP2A
NFKB2
ESR1
STAT1
TP53
WT1
EGFR

Recall – we should not over-interpret the role of individual edges!



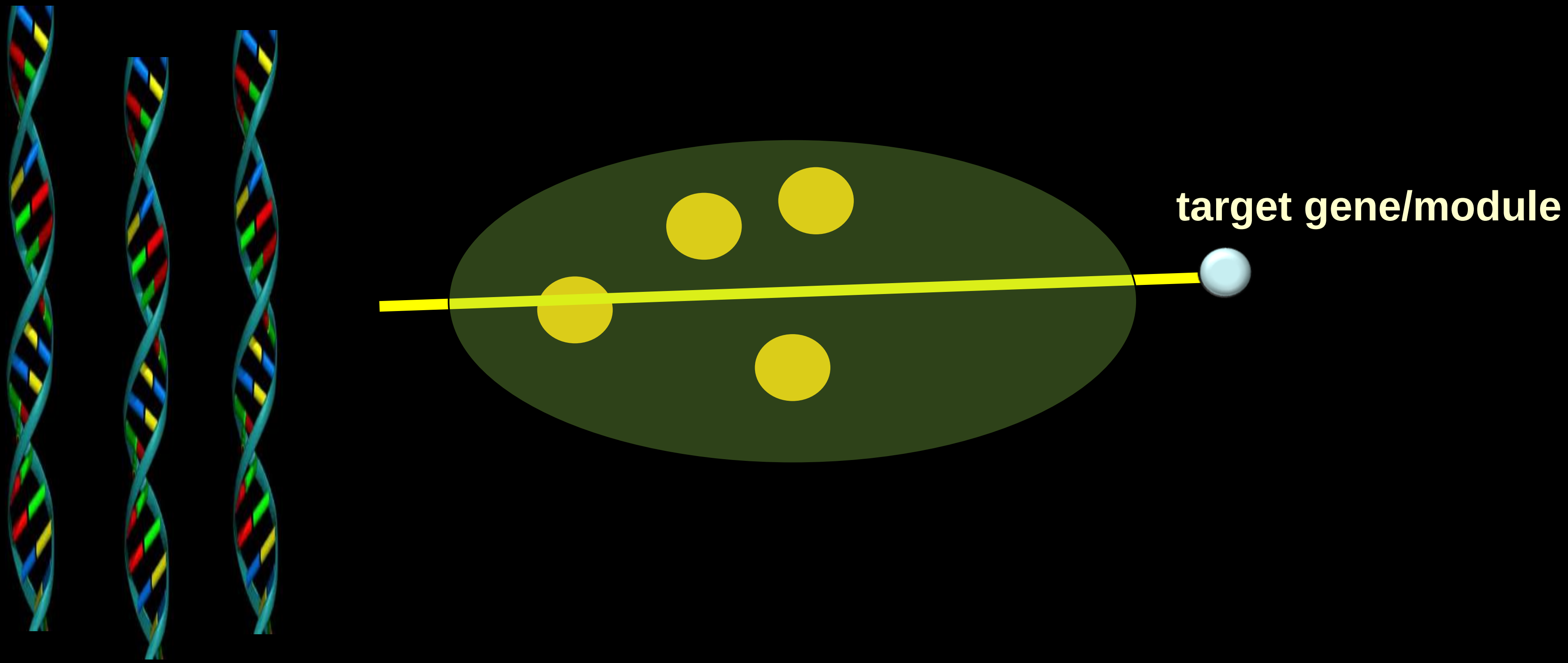
The Lute Player, Hendrick Maertensz Sorgh (1610-1670), Rijksmuseum, Amsterdam
(*public domain*)



Dutch Interior 1, Joan Miro' (1893–1983)
Museum of Modern Art, New York

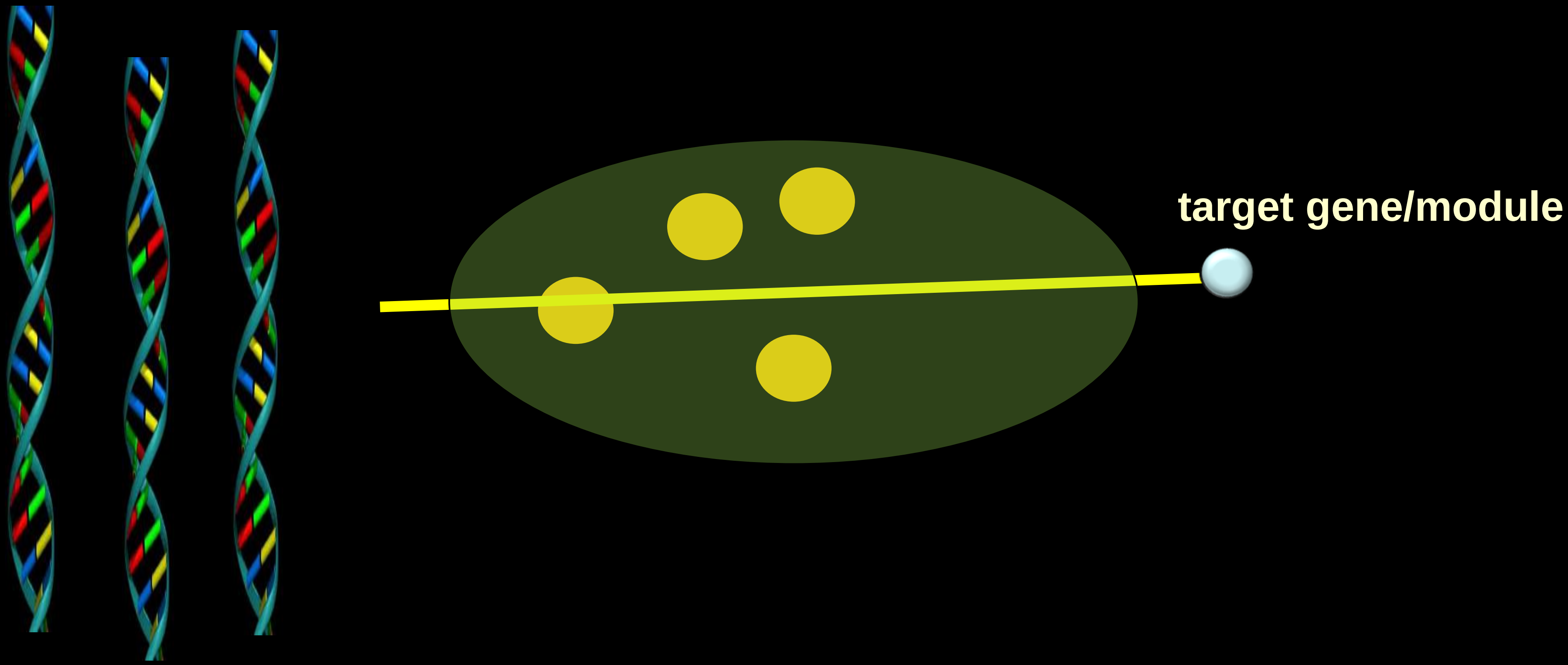
Cancer Cases
CNV data

Cancer Cases
Gene expression data



Kim et al. PloS CB 2011

Which pathways connect genotype to target gene ?

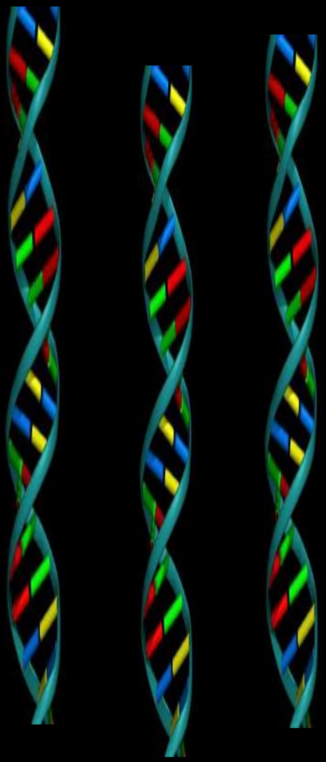


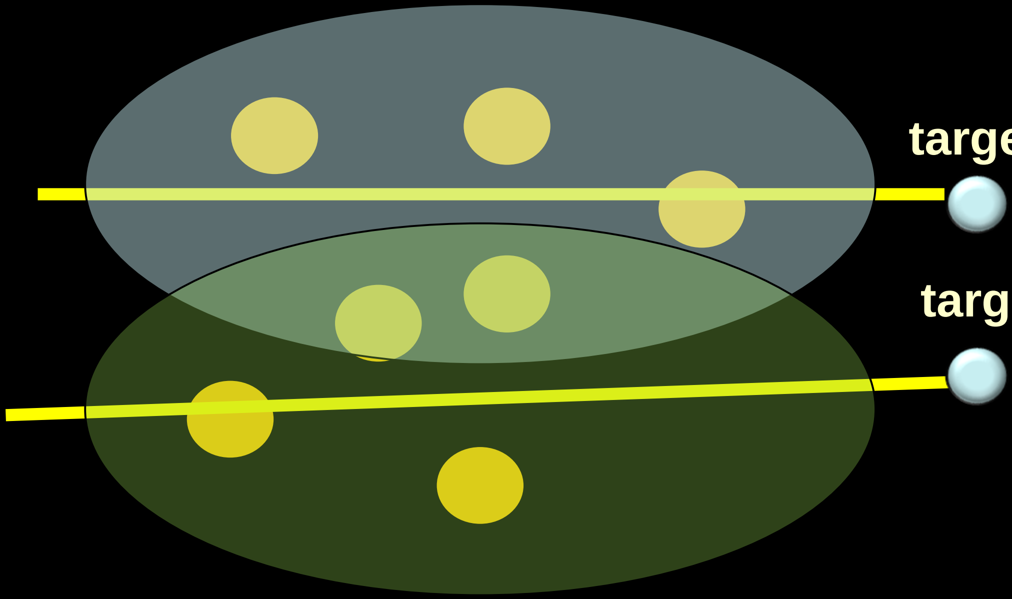
Kim et al. PloS CB 2011

Are there common functional pathways?

Cancer Cases
CNV data

Cancer Cases
Gene expression data





target gene/module

target gene/module

Kim et al. PloS CB 2011

Gene Hubs

MYC(110)	E2F1(88)	E2F4(43)	CREBBP(34)	GRB2(27)	SP3(26)	ESR1(25)
TFAP2A(25)	NFKB1(23)	MYB(22)	JUN(22)	E2F2(22)	RELA(21)	AR(21)
SP1(20)	RPS27A(20)	MAPK3(19)	POU5F1(17)	HIF1A(16)	PPARA(15)	CDC42(15)
UBA52(13)	CDK7(13)	YBX1(13)	YWHAZ(12)	CEBPB(12)	POU2F1(12)	UBE2I(11)
SMAD3(11)	TAL1(11)					

Pathway Hubs

GO biological process	#
cell cycle arrest	10
epidermal growth factor receptor signaling pathway	9
negative regulation of cell growth	9
Ras protein signal transduction	9
regulation of sequestering of triglyceride	8
cell proliferation	7
nuclear mRNA splicing, via spliceosome	7
regulation of cholesterol storage	7
nucleotide-excision repair	7
RNA elongation from RNA polymerase II promoter	7
insulin receptor signaling pathway	6
transcription initiation from RNA polymerase II promoter	6
N-terminal peptidyl-lysine acetylation	5
phosphoinositide-mediated signaling	5
positive regulation of lipid storage	4
positive regulation of specific transcription from RNA polymerase II promoter	3
positive regulation of epithelial cell proliferation	3
base-excision repair	2
negative regulation of hydrolase activity	2
gland development	2
positive regulation of MAP kinase activity	2
regulation of nitric-oxide synthase activity	2
estrogen receptor signaling pathway	2
regulation of receptor biosynthetic process	2
response to organic substance	2
JAK-STAT cascade	2
regulation of transforming growth factor-beta2 production	2
G1/S transition of mitotic cell cycle	2
SMAD protein nuclear translocation	2

Driving Copy number aberrations

ABCA1	ACP1	ADCY8	AGA	AHR	AKAP6	AKAP9
AKT1	ANXA11	ANXA2	APP	ARHGAP11A	ARHGAP29	ATR
BUB3	CAD	CAMK2G	CCNC	CDC2	CDC5L	→ CDKN2A
CEBPA	CEP70	CFH	CHUK	COBL	CRMP1	CSF2
CSNK2A1	CUL1	DARC	DDX56	DIAPH3	DLC1	EFNA5
→ EGFR	EIF2B1	EIF3A	EIF3B	EIF3F	ELMO1	EPB41
ERBB4	ERCC6	FAS	FER	FHL2	→ GBAS	GBE1
GSTK1	HEATR1	HSDL2	IFNA4	ILK	ITGB3BP	KITLG
LMO7	MAP2K4	MCM7	MED10	MON2	MRLC2	MS4A1
NDUFA4	NDUFB8	NRXN1	NUP205	NUPL1	ORC5L	PARP1
PCDH7	POLR1A	POLR2J	POLR3A	POLR3B	POM121	PPIA
PRIM1	PRKAB1	PRKCA	PSAP	PSMA1	PSMA4	PSMA5
PSMB1	PSMC3	PSMC6	→ PTEN	PTK2B	PTPRD	PTPRJ
PTPRK	RAI14	→ RB1	RBMX	RBPMS	REL	RGL1
RHOBTB2	RPL10	RPL10L	RPS17	SEC61A2	SF3B4	SFRS2
SFRS3	SGCB	SLC25A4	SLC27A2	SNRPB2	SPTA1	STXBP6
SYNGR1	TAF2	TERF2IP	THBS1	TOP1	→ TP53	TRIP13
TSSC1	U2AF2	UBE3A	USF2	VAV3	VDAC2	VIM
VWF	ZNF107					

Design details under the hood

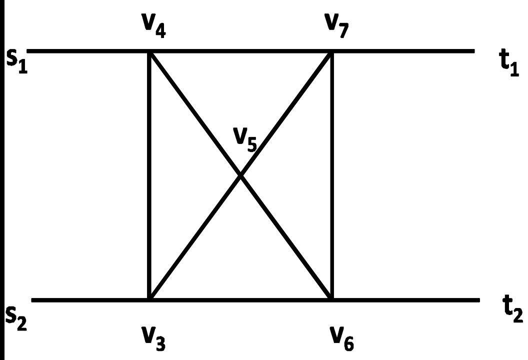
- Current flow reduces to solving a set of linear equations (Kirchhoff's laws)

  Caveat: We had to solving a linear system with 20,000 variables thousands of times for permutation test required some care

- Many biological interactions are directional. This can be taken care by solving linear program with corresponding constraints - Caveat: the network is to big for solving thousands of linear programs

- Null model and p-value estimations

Kim, Wuchty, Przytycka – *PloS Comp Bio 2011*
Kim, Przytycki, Wuchty, Przytycka – *Phys. Bio.* 2011



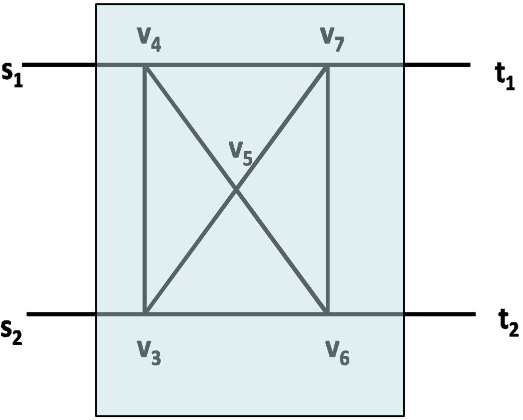
(a)

$$\begin{array}{|c|c|c|c|c|c|c|c|c|c|c|r|c|}
\hline
 & & & & & & & & & & & & C \\
\hline
1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & s_1{=}v_1 & 1/2 \\
\hline
0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & s_2{=}v_2 & 1/2 \\
\hline
0 & -1 & 4 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & v_3 & 0 \\
\hline
-1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & v_4 & 0 \\
\hline
0 & 0 & -1 & -1 & 4 & -1 & -1 & 0 & 0 & 0 & 0 & v_5 & 0 \\
\hline
0 & 0 & -1 & 0 & -1 & 4 & -1 & -1 & 0 & 0 & 0 & v_6 & 0 \\
\hline
0 & 0 & 0 & -1 & -1 & -1 & 4 & 0 & -1 & 0 & 0 & v_7 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 1 & 0 & t_1{=}v_8 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 1 & t_2{=}v_9 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & l_{t_1} & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & l_{t_2} & 0 \\
\hline
\end{array}$$

(b)

Kim, Przytycki, Wuchty, Przytycka – *Phys. Bio.* 2011

Rate limiting step inverting many matrices but all having common dense sub-matrix ….



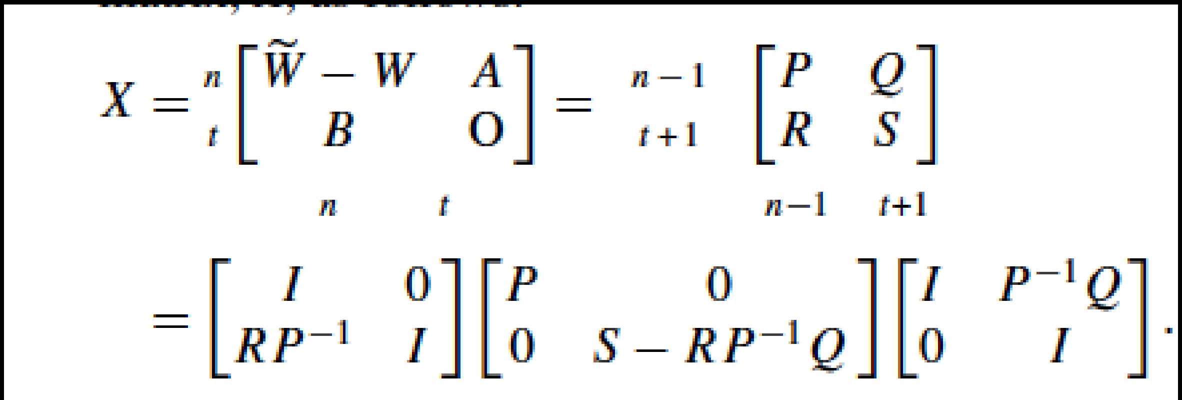
(a)

$$
\begin{array}{ccccccccccc|l}
1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & s_1=v_1 \\
0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & s_2=v_2 \\
0 & -1 & 4 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & v_3 \\
-1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & v_4 \\
0 & 0 & -1 & -1 & 4 & -1 & -1 & 0 & 0 & 0 & 0 & v_5 \\
0 & 0 & -1 & 0 & -1 & 4 & -1 & -1 & 0 & 0 & 0 & v_6 \\
0 & 0 & 0 & -1 & -1 & -1 & 4 & 0 & -1 & 0 & 0 & v_7 \\
0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 1 & 0 & t_1=v_8 \\
0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 1 & t_2=v_9 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & I_{t_1} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & I_{t_2}
\end{array}
\qquad
\begin{array}{c}
c \\ \hline
1/2 \\ 1/2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0
\end{array}
$$

(b)

Kim, Przytycki, Wuchty, Przytycka – *Phys. Bio.* 2011

Schur decomposition to minimize total cost of matrix inversions



$$X = \begin{array}{cc} \begin{matrix} n \\ t \end{matrix} & \begin{bmatrix} \widetilde{W} - W & A \\ B & O \end{bmatrix} \\ & \begin{matrix} n & \quad t \end{matrix} \end{array} = \begin{array}{cc} \begin{matrix} n-1 \\ t+1 \end{matrix} & \begin{bmatrix} P & Q \\ R & S \end{bmatrix} \\ & \begin{matrix} n-1 & t+1 \end{matrix} \end{array}$$

$$= \begin{bmatrix} I & 0 \\ RP^{-1} & I \end{bmatrix} \begin{bmatrix} P & 0 \\ 0 & S - RP^{-1}Q \end{bmatrix} \begin{bmatrix} I & P^{-1}Q \\ 0 & I \end{bmatrix}.$$

Note that the dense submatirx representing the network is common for all instances of the flow problem

Summary

Optimum connection approach

- **Shortest Path**
- **Steiner tree**

Return smaller number of genes easier to analyze from the perspective of individual genes.
More strongly depends of quality of network

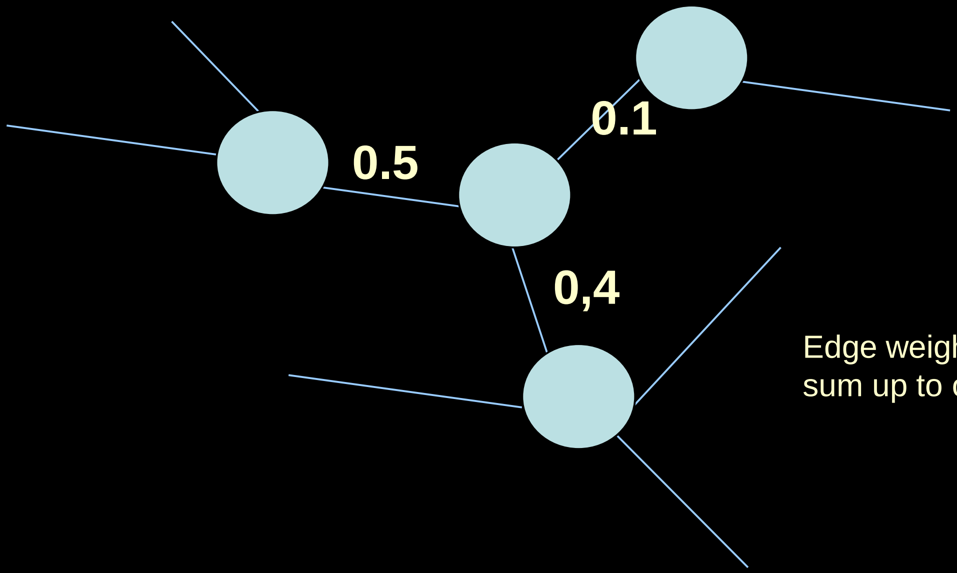
Information flow/ diffusion approach

- **Current Flow**
- **Hot Net**

More focused on group of genes and gene modules

Current Flow versus Random Walk

Current flow is equivalent (with appropriate edge weights) to the random walk: Starting at a given node move to an adjacent node with probability provided by edge weight, what is probability of ending at a terminal node starting at a given start node?



Edge weights around each node need to sum up to one

The equivalency is lost if we restrict the number of steps, loose information at each step etc.